SIGNAL PROCESSING FOR REAL-TIME DNA MICROARRAYS

H. Vikalo[1], *B. Hassibi*[2], *and A. Hassibi*[1]

[1]Department of Electrical and Computer Engineering, University of Texas, Austin, TX
[2]Department of Electrical Engineering, California Institute of Technology, Pasadena, CA

ABSTRACT

In conventional fluorescent-based microarrays, data is acquired after the completion of the hybridization phase. In this phase the target analytes (i.e., DNA fragments) bind to the capturing probes on the array and supposedly reach a steady state. Accordingly, microarray experiments essentially provide only a single, steady-state data point of the hybridization process. On the other hand, a novel technique (i.e., real-time microarrays) capable of recording the kinetics of hybridization in fluorescent-based microarrays has recently been proposed in [1]. The richness of the information obtained therein promises higher signal-to-noise ratio, smaller estimation error, and broader assay detection dynamic range compared to the conventional microarrays. In the current paper, we model the kinetics of the hybridization process measured by the real-time microarrays, and develop techniques for estimating the amounts of analytes present therein.

1. INTRODUCTION

A DNA microarray [2]-[4] is an affinity-based biosensor where the binding is based on hybridization, a chemical processes in which single DNA strands specifically bind to each other creating structures in a lower energy state. DNA microarrays are primarily used to measure gene expression levels, i.e., to quantify the process of transcription of DNA data into messenger RNA molecules (mRNA). The information transcribed into mRNA is further translated to proteins, the molecules that perform most of the functions in cells. Therefore, by measuring gene expression levels, researchers may be able to infer critical information about functionality of the cells or the whole organism.

Today, the sensitivity, dynamic range, and resolution of the DNA microarrays is limited by shot-noise, cross-hybridization, saturation, probe density variations, as well as several other sources of noise and systematic errors in the detection procedure. The number of hybridized molecules varies due to the probabilistic nature of the hybridization. It has been observed that these variations are very similar to shot-noise at high expression levels, yet more complex at low expression levels where the cross-hybridization becomes the dominating limiting factor of the signal strength [5]. Probe density variation further contribute to the uncertainty of the measurements. Additionally, saturation (which occurs when there are many more target molecules than the probe molecules in the corresponding spots) limits the achievable dynamic range.

Acquiring larger amounts of useful data (e.g., observing the entire hybridization process) would improve the SNR and the performance of microarrays. However, conventional fluorescent-based DNA microarrays are incapable of providing such additional data. There, the measured signal emanates from the fluorescently labeled target molecules which have hybridized to the probes at the surface of the microarray. Typically, the detection of the captured targets is carried out by scanning and/or various other imaging techniques after the hybridization step is completed and the solution is washed away. The reason for this is simple: a large concentration of floating (e.g., unbounded) labeled targets in the hybridization solution may overwhelm the specific signal emanating from the captured targets. Hence, conventional microarrays typically do not allow the presence of the solution during the fluorescent and reporter intensity measurements.

Recently, we have developed a novel *real-time microarray* (RT-μArray) system, capable of evaluating the abundance of multiple targets in a sample by performing real-time detection of the target-probe binding events [1]. This system samples fluorescent signals emanating from the probes capturing quencher-labeled targets in the solution and thus does not require any washing step. The RT-μArray systems may employ various time averaging schemes to suppress the Poisson noise and fluctuation of the target bindings. Due to these advantages, the RT-μArrays achieve higher SNR, potentially significantly smaller estimation error, and broader detection dynamic range compared to the conventional microarrays. The paradigm shift in data acquisition, from measuring a single steady-state data point in the conventional microarrays to obtaining full hybridization kinetics in the RT-μArray systems, requires novel detection algorithms. These need to be preceded by the development of probabilistic models of the hybridization process. There are relatively few attempts on modeling the kinetics of hybridization, and consec-

This work was supported in part by a Grubstake Award from California Institute of Technology, a grant from the David and Lucille Packard Foundation, and by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech.

utive experimental verification of those models. Examples include the real-time study of hybridization with optical wave guides in [6], and the study of the hybridization process in a fluorescence-based system with a single surface-bound probe and a single target in [7].

2. PROBABILISTIC MODEL

For the models developed in this paper, we assume that the hybridization in the microarrays under consideration is reaction-rate limited, rather than diffusion-limited. Assume that the hybridization process starts at $t = 0$, and consider discrete time intervals of length Δt. Consider the change in the number of bound target molecules during the time interval $(i\Delta t, (i+1)\Delta t)$. We can write

$$n_b(i+1) - n_b(i) = [n_t - n_b(i)]p_b(i)\Delta t - n_b(i)p_r(i)\Delta t,$$

where n_t denotes the total number of target molecules, $n_b(i)$ and $n_b(i+1)$ are the numbers of bound target molecules at $t = i\Delta t$ and $t = (i+1)\Delta t$, respectively, and where $p_b(i)$ and $p_r(i)$ denote the probabilities of a target molecule binding to and releasing from a capturing probe during the i^{th} time interval, respectively. Hence,

$$\frac{n_b(i+1) - n_b(i)}{\Delta t} = [n_t - n_b(i)]p_b(i) - n_b(i)p_r(i). \quad (1)$$

It is reasonable to assume that the probability of the target release does not change between time intervals, i.e., $p_r(i) = p_r$, for all i. On the other hand, the probability of forming a target-probe pair depends on the availability of the probes on the surface of the array. If we denote the number of probes in a spot by n_p, then we can model this probability as

$$p_b(i) = \left(1 - \frac{n_b(i)}{n_p}\right) p_b = \frac{n_p - n_b(i)}{n_p} p_b, \quad (2)$$

where p_b denotes the probability of forming a target-probe pair assuming an unlimited abundance of probes.

By combining (1) and (2) and letting $\Delta t \to 0$, we arrive to

$$\begin{aligned} \frac{dn_b}{dt} &= (n_t - n_b)\frac{n_p - n_b}{n_p} p_b - n_b p_r \\ &= n_t p_b - \left[(1 + \frac{n_t}{n_p})p_b + p_r\right] n_b + \frac{p_b}{n_p} n_b^2. \end{aligned} \quad (3)$$

Note that in (3), only $n_b = n_b(t)$, while all other quantities are constant parameters, albeit unknown.

Before proceeding any further, we will find it useful to denote

$$\alpha = (1 + \frac{n_t}{n_p})p_b + p_r, \quad \beta = n_t p_b, \quad \gamma = \frac{p_b}{n_p}. \quad (4)$$

Using (4), we can write (3) as

$$\frac{dn_b}{dt} = \beta - \alpha n_b + \gamma n_b^2 = \gamma(n_b - \lambda_1)(n_b - \lambda_2), \quad (5)$$

where λ_1 and λ_2 are introduced for convenience and are given by

$$\begin{aligned} \lambda_{1,2} &= \frac{n_p}{2}\left(\frac{p_r}{p_b} + 1 + \frac{n_t}{n_p}\right) \\ &\pm \frac{n_p}{2}\sqrt{\left(\frac{n_t}{n_p} - 1\right)^2 + \left(\frac{p_r}{p_b} + 1\right)^2 + 2\frac{n_t p_r}{n_p p_b} - 1}. \end{aligned}$$

Note that $\gamma = \beta/(\lambda_1\lambda_2)$. The solution to (5) is found as

$$n_b(t) = \lambda_1 + \frac{\lambda_1(\lambda_1 - \lambda_2)}{\lambda_2 e^{\beta(\frac{1}{\lambda_1} - \frac{1}{\lambda_2})t} - \lambda_1}. \quad (6)$$

From (5) (or (6)), it follows that

$$\beta = n_t p_b = \left.\frac{dy_b}{dt}\right|_{t=0}. \quad (7)$$

Therefore, the slope of the hybridization curve at $t = 0$ contains information about the amount of the target of interest. [Note that we may need to perform a calibration experiment to obtain p_b.] Estimating the amount of targets from the early stage of hybridization also alleviates the effect of saturation. In particular, since we do not wait for the steady-state of the reaction, we potentially enable a much broader dynamic range than that of conventional microarrays. This also implies potentially much faster detection than in conventional microarrays (minutes, compared to hours).

2.1. Estimating parameters of the model

Ultimately, by observing the hybridization process, we would like to obtain n_t, the number of target molecules. In addition, to fully characterize the hybridization process (including the computation of the reaction rate), we also need to find the parameters p_b , p_r, and n_p. However, we do not have direct access to $n_b(t)$ in (6), but rather to $y_b(t) = kn_b(t)$, where k denotes a transduction coefficient. In particular, we observe

$$y_b(t) = \lambda_1^* + \frac{\lambda_1^*(\lambda_1^* - \lambda_2^*)}{\lambda_2^* e^{\beta^*(\frac{1}{\lambda_1^*} - \frac{1}{\lambda_2^*})t} - \lambda_1^*}, \quad (8)$$

where

$$\lambda_1^* = k\lambda_1, \lambda_2^* = k\lambda_2, \text{ and } \beta^* = k\beta.$$

For convenience, we also introduce

$$\gamma^* = \frac{\beta^*}{\lambda_1^*\lambda_2^*} = \frac{\gamma}{k}, \text{ and } \alpha^* = \gamma^*(\lambda_1^* + \lambda_2^*) = \alpha. \quad (9)$$

From (8), it follows that

$$\beta^* = \left.\frac{dy_b}{dt}\right|_{t=0}. \quad (10)$$

Assume, without a loss of generality, that λ_1^* is the smaller and λ_2^* the larger of the two, i.e., $\lambda_1^* = \min(\lambda_1^*, \lambda_2^*)$ and $\lambda_2^* = \max(\lambda_1^*, \lambda_2^*)$. From (8), we find the steady-state of $y_b(t)$,

$$\lambda_1^* = \lim_{t\to\infty} y_b(t). \quad (11)$$

So, from (10) and (11) we can determine β^* and λ_1^*, two out of the three parameters in (8). To find the remaining one, λ_2^*, one needs to fit the curve (8) to the acquired data.

Having determined λ_1^*, λ_2^*, and β^*, we use (9) to obtain α^* and γ^*. Then, we may attempt to use (4) to obtain p_b, p_r, n_p, and n_t from α^*, β^*, and γ^*. However, (4) provides only 3 equations while there are 4 unknowns that need to be determined. Therefore, we need at least 2 different experiments to find all of the desired parameters. Assume that the arrays and the conditions in the two experiments are the same except for the target amounts applied. Denote the target amounts by n_{t_1} and n_{t_2}; on the other hand, it is reasonable to assume that p_b and p_r remain the same in the two experiments. Let the first experiment yield α_1^*, β_1^*, and γ_1^*, and the second one yield α_2^*, β_2^*, and γ_2^*, where $\gamma_2^* = \gamma_1^*$. Then it can be shown that

$$p_b = \frac{\beta_1^*\gamma_1^* - \beta_2^*\gamma_2^*}{\alpha_1^* - \alpha_2^*}, \quad (12)$$

and

$$p_r = \alpha_1^* - p_b - \frac{\beta_1^*\gamma_1^*}{p_b}. \quad (13)$$

Moreover,

$$n_p = \frac{p_b}{k\gamma_1^*}, \quad (14)$$

and

$$n_{t_1} = \frac{\beta_1^*\gamma_1^*}{p_b^2} n_p, \quad n_{t_2} = \frac{\beta_2^*\gamma_2^*}{p_b^2} n_p. \quad (15)$$

The following comments are in order. First, note that in (13)-(14) only the data obtained from one of the experiments (i.e., α_1^*, β_1^*, and γ_1^*) are used for the parameter estimation. As an alternative, we could repeat (13)-(14) using α_2^*, β_2^*, and γ_2^*, and then find p_r and n_p as the averages of their respective estimates. On another note, quantities (14)-(15) are known within the transduction coefficient k, where

$$k = \frac{y_b(0)}{n_p}.$$

To find k and thus unambiguously quantify n_p, n_{t_1}, and n_{t_2}, we need to perform a calibration experiment (i.e., an experiment with a known amount of targets n_t).

3. CANCELING CROSS-HYBRIDIZATION

Expression (3) describes the change in the amount of target molecules, n_b, captured by the probes in a single probe spot of the microarray. Similar equations hold for other spots and other targets. Moreover, (3) can be extended to model kinetics of both hybridization and cross-hybridization (i.e., non-specific binding). For instance, if we assume that the signal measured by a particular probe spot consists of a hybridization and a cross-hybridization component, they can be described by the following system of coupled differential equations,

$$\begin{aligned}\frac{dn_{b,h}}{dt} &= (n_h - n_{b,h})\frac{n_p - n_{b,h} - n_{b,c}}{n_p} p_h - n_{b,h}p_{r,h},\\ \frac{dn_{b,c}}{dt} &= (n_c - n_{b,c})\frac{n_p - n_{b,h} - n_{b,c}}{n_p} p_c - n_{b,c}p_{r,c},\end{aligned}$$

where $n_{b,h}$ and $n_{b,c}$ denote the number of specific and non-specific targets bound to probes, n_h and n_c denote the total number of specific and non-specific targets, and where p_h and p_c denote the probabilities of forming specific and non-specific target-probe pairs given an unlimited abundance of the probe molecules while $p_{r,h}$ and $p_{r,c}$ denote the probabilities of breaking those pairs, respectively.

Focusing on the early phase of the hybridization process and its reaction rate opens up the possibility of suppressing cross-hybridization. When a single target analyte is present, the number of available probe molecules, or equivalently the light intensity of a probe spot, decays exponentially with time as $Ce^{-\alpha t}$, where α is as in (4), and where C is determined from β, γ, and the initial light intensity of the probe spot. If, in addition to hybridization of the target of interest, a number of other targets cross-hybridize to the same probe spot, the light intensity of the probe spot will decay as the sum of several exponentials,

$$I(t) = \sum_{k=0}^{K} C_k e^{-\alpha_k t}, \quad (16)$$

where index $k = 0$ corresponds to the desired target, and $k = 1, \ldots, K$ correspond to the cross-hybridizing analytes. The reaction rates for the different analytes differ due to different numbers of analytes, binding probabilities, etc. (we omit explicit expressions for brevity). Therefore, if we can estimate the reaction rates from (16), we should be able to determine the number of molecules for each of the analytes binding to the spot.

The RT-uArray system samples the signal (i.e., the light intensity) of the probe spots at certain time intervals (multiples of Δ, say) and thus obtains the sequence

$$y_n = I(n\Delta) + v(n\Delta) = \sum_{k=0}^{K} C_k e^{-n\Delta\alpha_k} + v(n\Delta),$$

for $n = 0, 1, \ldots, T$, where T is the total number of samples, and $v(t)$ represents the measurement noise. Defining $u_k = e^{-\Delta\alpha_k}$, we may write

$$y_n = \sum_{k=0}^{K} C_k u_k^n + v(n), \tag{17}$$

The goal is to (i) determine the value of K (i.e., how many analytes are binding to the probe spot), (ii) estimate the values of the pairs $\{C_k, u_k\}$ for all $k = 1, \ldots, K-1$, and (iii) determine the number of each analyte.

The problem of determining the number of exponential signals in noisy measurements, and estimating the individual rates, is a classical one in signal processing and is generally referred to as system identification. (There are a multitude of books and papers on this subject.) The basic idea is that, when y_n is the sum of K exponentials, it satisfies a Kth order recurrence equation

$$y_n + h_1 y_{n-1} + \cdots + h_{K-1} y_{n-K+1} + h_K y_{n-K} = 0.$$

Furthermore, the u_k are the roots of the polynomial

$$H(z) = z^K + h_1 z^{K-1} + \cdots + h_{K-1} z + h_K.$$

In practice, since one observes a noisy signal, one first uses the measurements to form the so-called Hankel matrix,

$$\begin{bmatrix} y_{T/2} & y_{T/2-1} & \cdots & y_1 & y_0 \\ y_{T/2+1} & y_{T/2} & \cdots & y_2 & y_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_T & y_{T-1} & \cdots & y_{T/2+1} & y_{T/2} \end{bmatrix}.$$

When y_n is the sum of K exponentials, the above Hankel matrix has rank K, i.e., only K nonzero eigenvalues. When y_n is noisy, the standard practice is to compute the singular values of the Hankel matrix and estimate K as being the number of significant singular values.

Once K has been determined, one forms the $(T-K+1) \times (K+1)$ Hankel matrix

$$\begin{bmatrix} y_K & y_{K-1} & \cdots & y_1 & y_0 \\ y_{K+1} & y_K & \cdots & y_2 & y_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_T & y_{T-1} & \cdots & y_{T-K+1} & y_{T-K} \end{bmatrix} \tag{18}$$

and then identifies the vector $[h_1 \ \ldots \ h_K]$ with the smallest right singular vector of (18).

As mentioned earlier, the roots of $H(z)$ are the desired u_k, from which we determine the rates α_k and thereby the amounts of targets present. While the main idea was outlined above, we may use a variety of different techniques to find the u_k, including – but not limited to – total least squares, ESPRIT, Prony's method, etc. [See, e.g., [9], [10], and the references therein.]

The performance of one such algorithm is illustrate by simulations in Figure 1. In particular, we consider the so-called total least squares (TLS) algorithm (see, e.g., [9]) in the situation where two target analytes bind to the same probe spot – one due to hybridization, and the other due to cross-hybridization. Parameters of the system (probabilities of hybridization, cross-hybridization, release, etc.) are chosen so as to mimic realistic experimental scenarios. The probability of hybridization is assumed to be 5 times greater than the probability of cross-hybridization (i.e., $p_h/p_c = 5$). The number of hybridizing target is $n_h = 10^9$, while the number of cross-hybridizing molecules is varied. In Figure 1, we plot the relative mean-square error of estimating n_h (averaged over many realizations of noise) as a function of the ratio n_h/n_c. The simulation results indicate potentially successful suppression of cross-hybridization over 3 orders of magnitude of n_h/n_c.

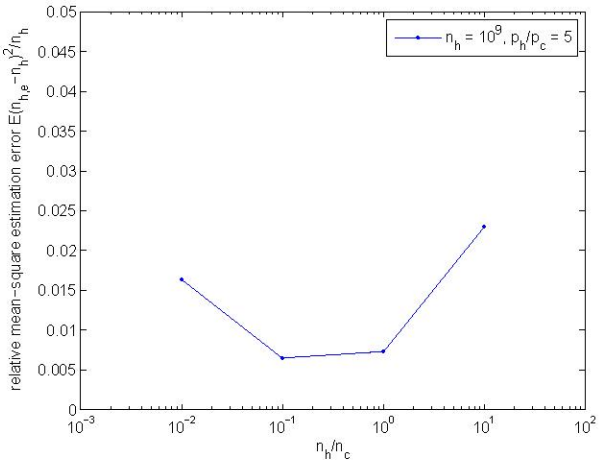

Fig. 1. *The relative mean-square error of estimating n_h (averaged over many realizations of noise) as a function of the ratio n_h/n_c, where $n_h = 10^9$ and $p_h/p_c = 5$.*

4. EXPERIMENTAL RESULTS

To test the validity of the proposed model and demonstrate the parameter estimation procedure, we designed and conducted two DNA microarray experiments. We designed custom 8×9 arrays containing 25mer probes printed with 3 different densities. The targets were mRNA Spikes purchased from Ambion, Inc., applied to the arrays with different concentrations. The concentrations used in the two experiments were 80ng/50μl and 16ng/50μl.

The signal measured in the first experiment, where 80ng of the target is applied to the array, is shown in Figure 2. The smooth line shown in the same figure represents the fit obtained according to (6). In the second ex-

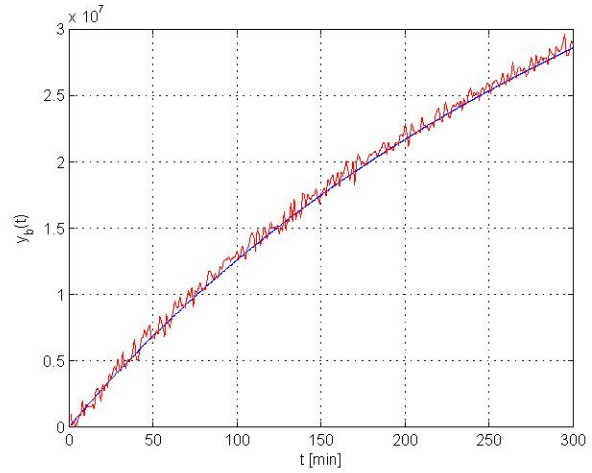

Fig. 2. *The measured signal from 80ng of Ambion Spike 3 applied to a microarray, and the fit according to (6).*

periment, 16ng of the target is applied to the array. The measured signal, and the corresponding fit obtained according to (6), are both shown in Figure 3.

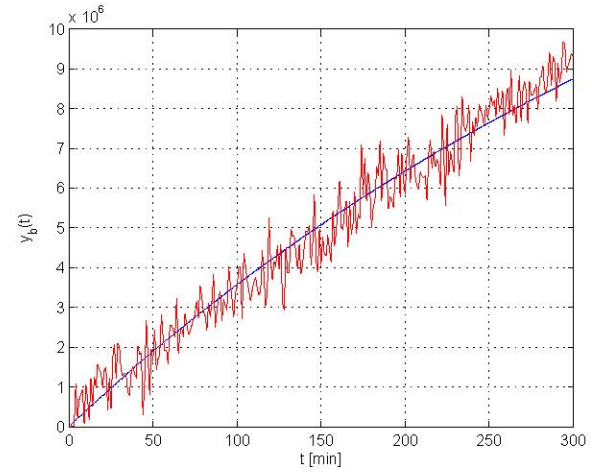

Fig. 3. *The measured signal from 16ng of Ambion Spike 3 applied to a microarray, and the fit according to (6).*

From the two experiments, we find that

$$n_{t_1}/n_{t_2} = \beta_1/\beta_2 = 3.75. \qquad (19)$$

Note that the above ratio is relatively close to its true value, $80/16 = 5$. Moreover, as explicitly shown in [8], using the data acquired in two experiments, we can estimate all of the parameters: n_{t_1}, n_{t_2}, n_p, p_b, and p_r.

5. SUMMARY AND CONCLUSIONS

To summarize, real-time microarrays acquire full kinetics of a hybridization process. We derived a probabilistic model of the hybridization process, proposed a simple estimation procedure based on the early phase of the hybridization, and showed how to cancel effects of cross-hybridization. The use of signal processing techniques may enable real-time microarrays to achieve higher signal-to-noise ratio and broader dynamic range than conventional microarrays.

6. REFERENCES

[1] A. Hassibi, H. Vikalo, and B. Hassibi, "Real-time microarrays," to be submitted to *PNAS*, 2007.

[2] M. Schena, *Microarray Analysis*, John Wiley & Sons, 2003.

[3] U. R. Mueller and D.V. Nicolau (Eds.), *Microarray Technology and Its Applications*, Springer, Berlin, Germany, 2005.

[4] W. Zhang and I. Shmulevich (Eds.), *Computational and Statistical Approaches to Genomics*, Kluwer Academic Publishers, 2002.

[5] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *PNAS*, October 29, 2002, 14031-14036.

[6] D. I. Stimpson et. al., "Real-time detection of DNA hybridization and melting on oligonucleotide arrays by using optical wave guides," *Proc. Natl. Acad. Sci. USA*, vol. 92, July 1995, 6379-6383.

[7] M. R. Henry, P. W. Stevens, J. Sun, and D. M. Kelso, "Real-time measurements of DNA hybridization on micropartcles with fluorescence resonance energy transfer," *Analyt. Biochem.*, no. 276, 1999, 204-214.

[8] H. Vikalo, B. Hassibi, M. Stojnic, and A. Hassibi, "Modeling the kinetics of hybridization in microarrays," *IEEE Intern. Workshop on Genomic Signal Process. and Statistics*, Tuusula, Finland, 2007.

[9] E. M. Dowling et. al., "Exponential Parameter Estimation in the Presence of Known Components and Noise," *IEEE Trans. on Antennas and Propag.*, vol. 42, no. 5, May 1994.

[10] A. J. van der Veen, E. F. Deprettere, and A. L. Swindlehurst, "Subspace Based Signal Analysis Using Singular Value Decomposition," *Proc. of the IEEE*, 81(9):1277-1308, September 1993.